

Conference Abstract

AI-Driven Solutions for Optimising Drawer Expansion and Storage Management of Entomology Collections

Qianqian Gu[‡], Ben Scott[‡], Geoff Martin[‡], Arianna Salili-James[‡], Laurence Livermore[‡], Peter Wing[‡], Vincent S. Smith[‡]

[‡] The Natural History Museum, London, United Kingdom

Corresponding author: Qianqian Gu (qianqian.gu@nhm.ac.uk)

Received: 15 Dec 2025 | Published: 14 Jan 2026

Citation: Gu Q, Scott B, Martin G, Salili-James A, Livermore L, Wing P, Smith VS (2026) AI-Driven Solutions for Optimising Drawer Expansion and Storage Management of Entomology Collections. Biodiversity Information Science and Standards 10: e182583. <https://doi.org/10.3897/biss.10.182583>

Abstract

The digitisation of natural history collections at scale raises a range of logistical and curatorial challenges. One concern is the physical expansion of storage infrastructure. As part of the Distributed System of Scientific Collections UK (Smith et al. 2022), the Natural History Museum London (NHM) will rehouse a significant proportion of its 36 million pinned entomological specimens (~130,000 drawers). Digitisation involves imaging every specimen and attaching a physical, unique, human- and machine-readable identifier (a small card label with a Data Matrix barcode). These barcodes must be readable from above, and their addition during digitisation can increase drawer occupancy if they exceed the existing footprint. Manual estimation of the footprint increase is impractical at this scale, given the range of circumstances associated with different specimens and drawers.

To address this, the NHM has developed an AI-driven approach to estimate specimen drawer expansion and support resource planning. The deep learning pipeline can automatically detect drawer objects, referred to here as classes, including specimens, labels, barcodes, notes, unit trays, and drawers, from high-resolution images. Our dataset includes 11,090 digitised pinned Coleoptera drawers from the Index Lot collection

(Natural History Museum 2014), representative of typical entomological collections. The AI pipeline supports both "pre-" and "post-" digitisation analysis by calculating the specimen bounding areas, handling overlaps, and estimating the net footprint change.

Our object detection models achieved high performance under real-world conditions, with a mean average precision (mAP^{*1}) of 85.45% across all classes. Barcode detection reached 86.63% mAP, while the standard unit tray detection and unit tray type classification model achieved 99.5% accuracy. (An example of the detection outputs is shown in Fig. 1)

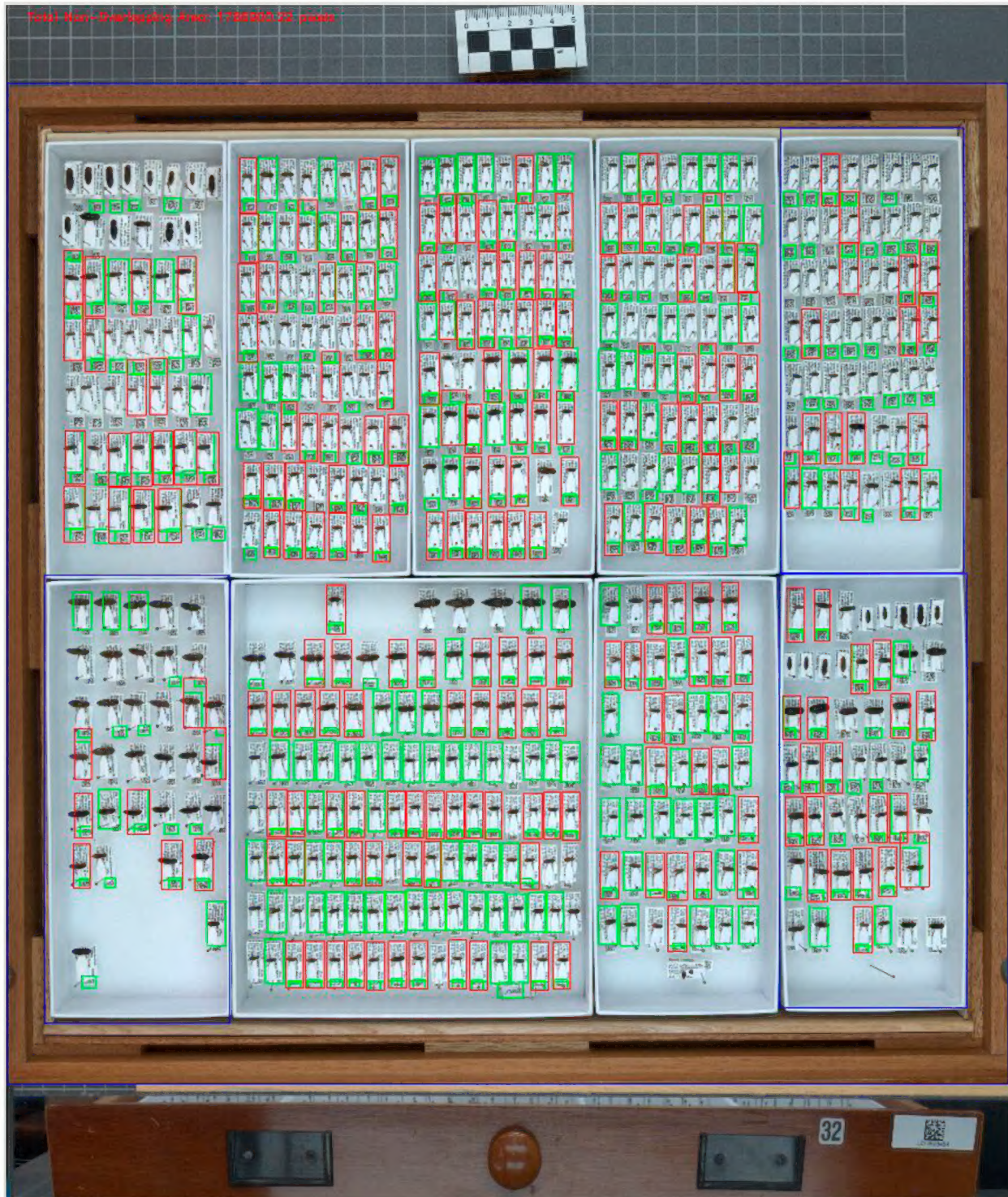


Figure 1.

Detection model outputs: green = specimen occupancies/barcodes; red = paired specimens/barcodes; blue = drawers/unit trays. Original drawer image © Trustees of the Natural History Museum, London, used with permission; model-generated overlays by the authors, licensed under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

Three different calculation methods were tested and evaluated to estimate the drawer expansion area, corresponding to three specimen expansion rates for specimens with Data Matrix barcodes attached.

- Rate 1 (Per-Drawer Area via Polygon Masks²): calculates expansion rates per drawer by measuring the total area occupied by specimens and their labels, ensuring overlaps are counted only once using polygon masks, as shown in Fig. 2.
- Rate 2 (Total-Based Expansion Rate): calculates expansion rates by averaging the occupancy areas of barcode-attached specimens across the entire dataset.
- Rate 3 (Per-Specimen Expansion Rate): averages the expansion rate per specimen using only bounding box³ areas. Every specimen contributes equally, regardless of its size, to the final value.

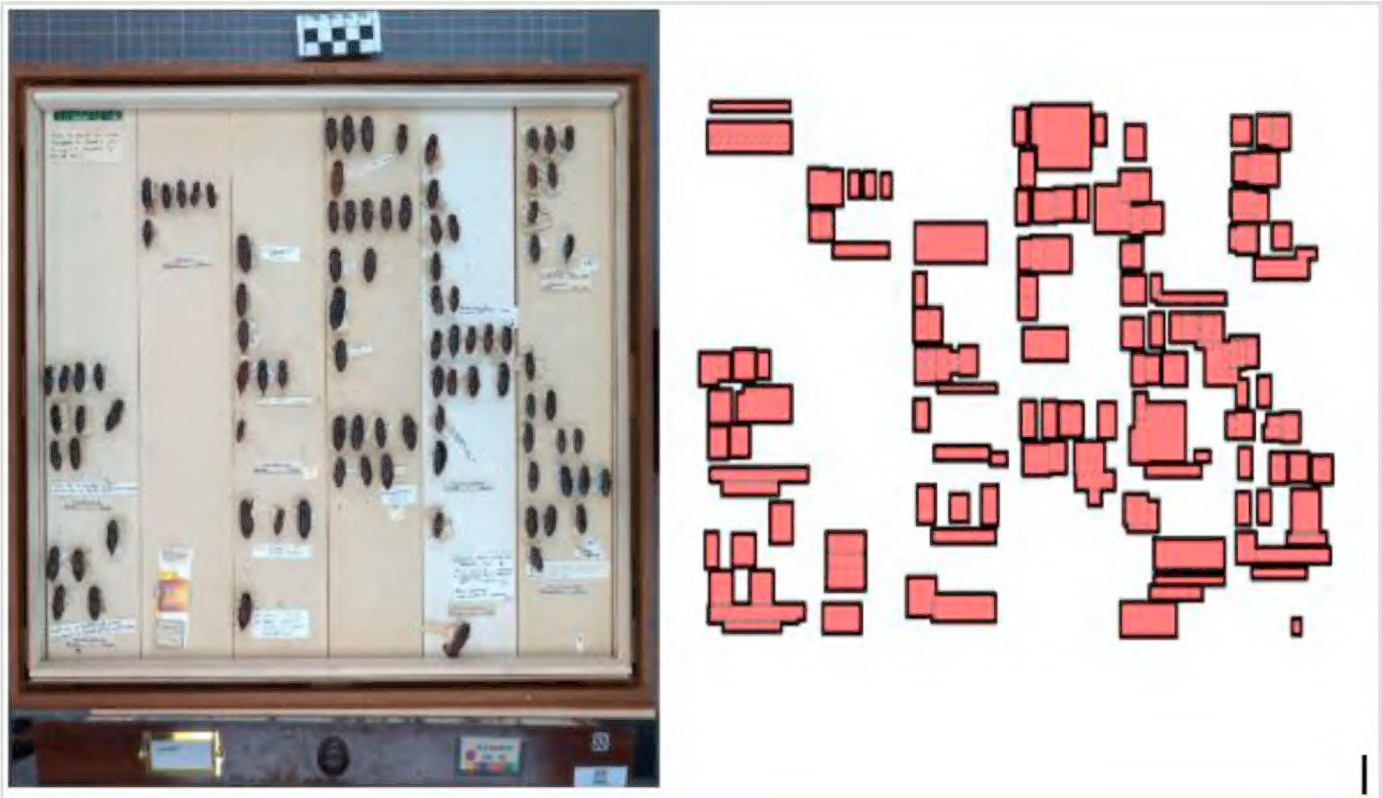


Figure 2.
Polygon masks overlaid on an Entomology drawer (left: image; right: model output). Original drawer image © Trustees of the Natural History Museum, London, used with permission; masks generated by the authors' model, licensed under [CC BY 4.0](#)

Table 1 compares different expansion rates. Both per-drawer Rate 1 and per-specimen Rate 3 have confidence intervals. Fig. 3 and Fig. 4 show their distributions: Rate 1 is right-skewed, indicating modest area increases from overlap, while Rate 3 is bimodal⁴, with a small peak near 0 (large specimens unchanged) and another around 0.3 (barcode additions on smaller specimens). The global total-based Rate 2 is the mean of the ratios without confidence intervals, providing a macro-level view.

In addition to estimating drawer expansion and supporting budget planning, this tool has been designed in a modular fashion, allowing individual components to be used independently at different stages of the digitisation and curation workflow. For example,

specific models can be used to flag missing labels or barcodes during digitisation, assist in tracking specimen relocation, and support downstream re-curation decisions. Importantly, the pipeline is designed to be reusable across other entomological collections, with all code to be made openly available alongside a forthcoming publication, promoting scalability within the community.

Table 1.
This table presents the average percentage increase in area across different expansion rate calculation methods, along with their 95% confidence intervals, full observed ranges, and pros and cons.

Name	Rate (%)	95% Confidence Interval	Pros	Cons
Specimen Expansion Rate 1 (Per Drawer)	12.25 (12.04~12.46)	Yes	Reflects real layout; overlap-aware	Layout-dependent; may understate impact when labels are tightly packed
Specimen Expansion Rate 2 (Global: Total based)	19.06	No	Best for total storage estimates; simple	Hides specimen-level variability; may be biased toward large specimens
Specimen Expansion Rate 3 (Per Specimen)	21.90 (21.80~22.00)	Yes	Captures specimen-level variation; supports CI and distributions	Noisy for small specimens; not ideal for total space estimation

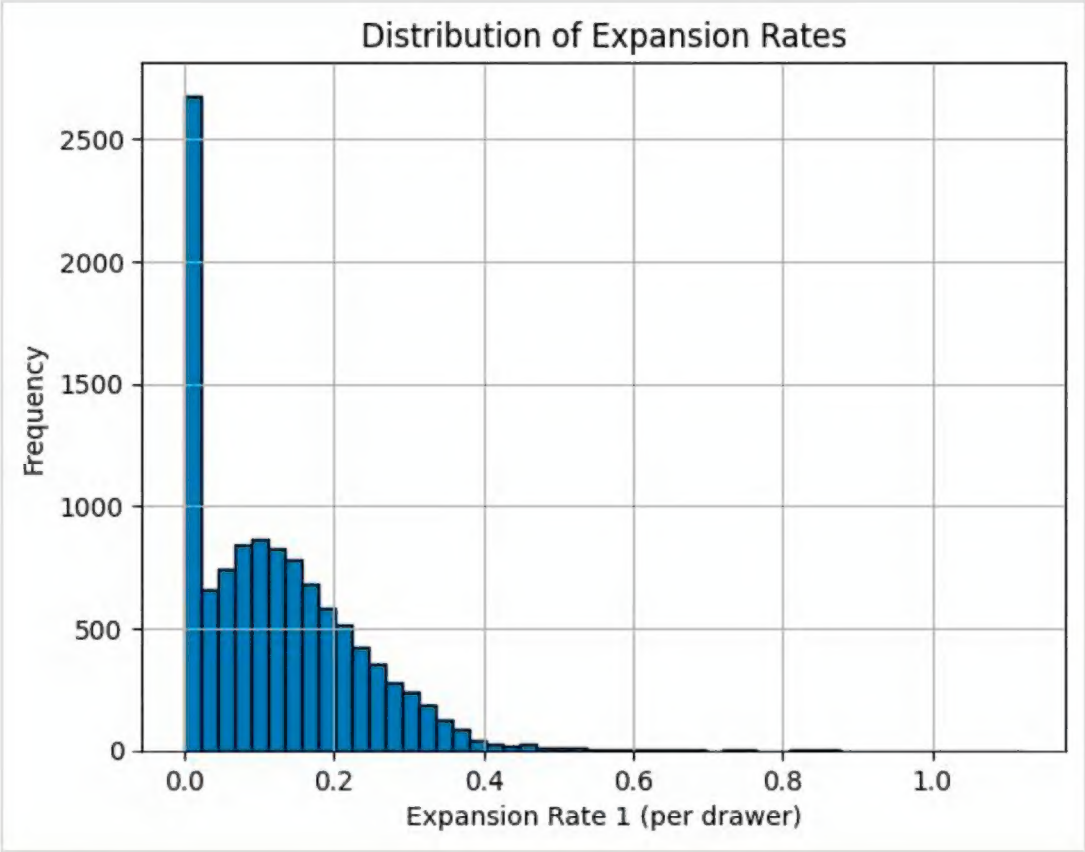


Figure 3.
The Expansion Rate 1 (per-drawer area via polygon masks) distribution shows a single tight right-skewed peak, with most drawers below 0.2. This reflects how per-drawer calculations preserve curators’ arrangement preferences but can reduce specimen visibility where items are densely placed, resulting in only modest area increases.

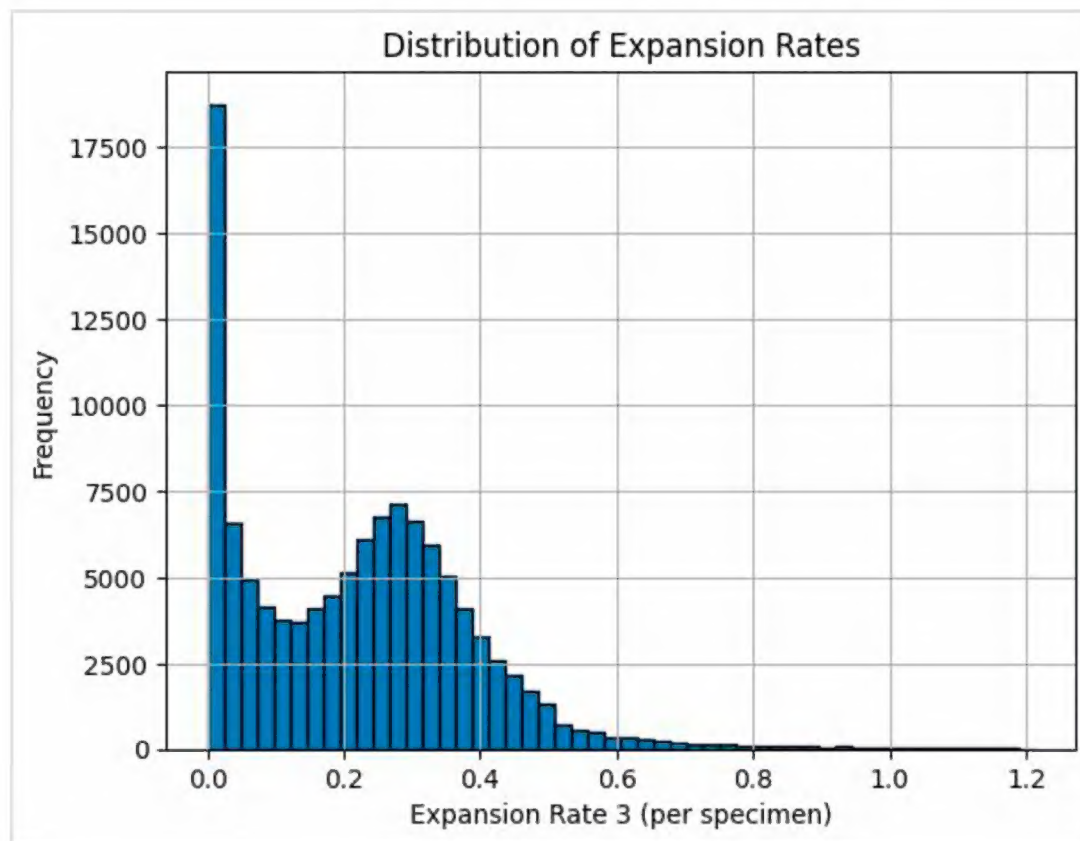


Figure 4.

The Expansion Rate 3 (per-specimen bounding box area) distribution is bimodal, reflecting strong specimen-level visibility, as each specimen contributes equally. A peak near 0 likely occurs because barcodes cause little change in large specimens, while the peak around 0.3 suggests barcode additions noticeably increased the area for many smaller specimens.

Our findings demonstrate that automated spatial analysis not only improves accuracy and speed in collection management but also lays the groundwork for predictive infrastructure modelling across large-scale digitisation efforts.

Keywords

digitisation, storage optimisation, computer vision, machine learning, natural history collections

Presenting author

Qianqian Gu

Presented at

Living Data 2025

Acknowledgements

The Digitisation Team, Natural History Museum, London

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Natural History Museum (2014) Index Lot collection Data set. Natural History Museum URL: <https://doi.org/10.5519/0073880>
- Smith VS, Hardy H, Wainwright T, et al. (2022) Harnessing the power of natural science collections: a blueprint for the UK. Natural History Museum <https://doi.org/10.5281/zenodo.6472239>

Endnotes

- *1 Mean Average Precision (mAP): a metric used to evaluate the performance of ranked retrieval and object detection models by averaging the precision scores at different recall levels.
- *2 Polygon Mask (in computer vision): a shape defined by a set of vertices used to identify and isolate a specific area of an image for processing.
- *3 Bounding Box: a rectangular outline that defines the position and extent of an object or region of interest.
- *4 Bimodal Distribution: a type of probability distribution with two distinct peaks, indicating two separate groups within a dataset.